DOCUMENT RESUME

ED 318 751 TM 014 767

AUTHOR Hinkle, Dennis E.; Winstead, Wayland H.
TITLE Bootstrap Methods: A Very Leisurely Look.
PUB DATE Apr 90
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Correlation; Error of Measurement; *Estimation (Mathematics); Factor Analysis; High Schools; High School Seniors; Monte Carlo Methods; National Surveys; Regression (Statistics)
IDENTIFIERS *Bootstrap Methods; High School and Beyond (NCES); Statistical Analysis System

ABSTRACT

The Bootstrap method, a computer-intensive statistical method of estimation, is illustrated using a simple and efficient Statistical Analysis System (SAS) routine. The utility of the method for generating unknown parameters, including standard errors for simple statistics, regression coefficients, discriminant function coefficients, and factor weights is also demonstrated. Data were drawn from the High School and Beyond Survey, a longitudinal study of high school sophomores and seniors. A sample of 500 (out of 28,240) seniors was selected. Using the SAS routine, 100, 200, and 500 Bootstrap samples were used to compute the Bootstrap estimate and its standard error for the following parameters: (1) means; (2) correlation coefficient; (3) simple regression coefficient; (4) multiple regression coefficient; (5) discriminant function coefficient; and (6) factor loadings. Research literature suggests that Bootstrap samples are used to estimate parameters and generate their standard errors and to develop non-parametric confidence intervals. The use of Bootstrap and Jackknife methods may have been limited by researchers' lack of familiarity with Monte Carlo methods or the misconception that such methods require enormous computer resources. One table lists the SAS routine, and five tables illustrate the method used. (SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Dennis E. Hinkle

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED318751

# Bootstrap Methods: A Very Leisurely Look

Dennis E. Hinkle
Wayland H. Winstead

Virginia Polytechnic Institute
and State University

Presented at the Annual Meeting of the American Educational Research Association, Boston, MA, April, 1990

TM 014767



BEST COPY AVAILABLE

# Bootstrap Methods:
# A Very Leisurely Look

**Dennis E. Hinkle**
**Wayland H. Winstead**

**Virginia Polytechnic Institute**
**and State University**

A common problem in the application of statistics to research problems in the social and behavioral sciences involves the estimation of some unknown population parameter, such as a mean, correlation coefficient, regression, discriminant,or factor weight, etc., from a sample statistic. The specific concern is related to how the sample statistic, calculated from random samples, varies from sample to sample (Diaconis & Efron, 1983; Efron & Gong, 1983; Efron & Tibshirani, 1986; Miller, 1974). This issue is referred to as "statistical accuracy." In determining the statistical accuracy of an estimate, i.e., developing confidence intervals for the parameter or testing hypotheses about the parameter, the traditional statistical approach relies on calculating the standard error of the estimate under the normality assumption. Often times, however, this assumption is tenuous for data collected in the social and behavioral sciences (Bickel & Freedman, 1981; Efron & Gong, 1983).

To a great extent, the reliance on using normal distribution theory in statistical estimation and hypothesis testing is a function of computational ease. Given the computation devices of the late 1800s and early 1900s, it is easy to understand why much of the research done in the area of statistical distribution theory focused on the distributional characteristics of both empirical and theoretical sampling distributions. However, with the advent of powerful computing resources in the late 1900s, it has been possible to validate the distributional characteristics of these sampling distributions as well as to

propose and to develop new computer intensive methods for estimating the unknown parameters (Diaconis & Efron, 1983).

The two most commonly cited computer intensive methods are the Jackknife (Tukey, 1958; Miller, 1974) and the Bootstrap (Efron, 1979). The Jackknife method was proposed by Tukey in 1958, based upon the research done by Quenouille in 1949 and 1956, as the "all purpose statistical tool" analogous to the Swiss Pocket Knife. The jackknife method is often referred to as the "leave one out" technique. To illustrate the method, consider a research study in which achievement scores are to be predicted from aptitude scores. Assuming n = 24, the jackknife method involves computing the regression coefficient from 23 out of the 24 observations. Twenty-four different coefficients are computed from 24 different samples by sequentially leaving out one paired observation. The estimated parameter, $\hat{\theta}$, is simply the mean of the observed sample coefficients; the standard error of this estimate, $s_{\hat{\theta}}$, is simply the standard deviation of the sample coefficients. Over the past thirty years, much research has been done on the statistical accuracy of the estimate using the Jackknife method (see Miller, 1974). These studies indicate that these estimates are unbiased for homoscedastic errors and robust against heteroscedasticity (Wu, 1985).

The Bootstrap method was developed by Efron in 1977 (Diaconis & Efron, 1983) as an alternative to the Jackknife method; the term comes from the expression "pulling yourself up by the bootstraps." In statistical terms, it refers to the fact that one available sample gives rise to many other samples. The process involves taking multiple samples from the original sample, using sampling with replacement. For the above research study, the Bootstrap method would involve randomly selecting (with replacement) 24 paired observations. For each sample, the regression coefficient would be computed, and, like the Jackknife method, the estimated parameter ($\hat{\theta}$) is simply the mean of the observed sample coefficients. Similarly, the standard error of this estimate ($s_{\hat{\theta}}$) is the

standard deviation of the sample coefficients. Like the Jackknife method, these estimates are unbiased for homoscedastic errors and robust against heteroscedasticity (Wu, 1985).

## Purpose Statement

Computer intensive statistical methods, including the Jackknife and Bootstrap methods, have been shown to be appropriate alternatives to traditional statistical methods that rely heavily upon normal distribution theory. While the application of these methods has been illustrated in a variety of research settings, their use has been somewhat limited. One reason may be researchers' lack of understanding and/or familiarly with Monte Carlo methods. A second reason may be the perceptions that such methods require enormous computer resources, either mainframe or micro.

The purpose of this paper is to illustrate the Bootstrap method using a simple and efficient SAS routine. The second purpose is to illustrate the method the utility of this method for generating standard errors for simple statistics (Efron & Gong, 1983), regression coefficients (Freedman, 1981), discriminant function coefficients (Fenwick, 1979), and factor weights (Jennrich & Clarkson, 1980; Chattergee, 1984). The data to be used for these illustrations were drawn from the High School and Beyond Survey, a longitudinal study of high school sophomores and seniors sponsored by the National Center for Educational Statistics (the data set is described in Hinkle, Wiersma and Jurs, 1988). A sample of 500 of the 28,240 seniors was selected and the data for the following variables were recorded. A second sample of 75 was selected from the original data set. The purpose of this second sample was pedagogical only. (Copies of these data sets can be obtained from the first author.)

**MATHACH** A 25-item mathematics achievement test with scores ranging from -8.33 to 25.

**VISUAL** A 16-item test of visualization in three dimensions, for determining how a three-dimensional object would look if its spatial position were changed; scores range from -4 to 16.

| | |
|---|---|
| **MOSAIC** | A 56-item test involving the detection of relationships in patterns of tiles; scores range from -28 to 56. |
| **GRADES** | Student's report of grades, recorded such that an 8 represents mostly A's and 1 represents below D. |
| **MATHGR** | Student's report of grades in mathematics courses mostly A's and B's is coded 1, otherwise 0. |
| **ALG1** | Enrollment in Algebra 1; 1 indicates that the student has taken the course, 0 indicates that the course was not taken. |
| **ALG2** | Enrollment in Algebra 2; 1 indicates that the student has taken the course, 0 indicates that the course was not taken. |
| **GEO** | Enrollment in Geometry; 1 indicates that the student has taken the course, 0 indicates that the course was not taken. |
| **TRIG** | Enrollment in Trigonometry; 1 indicates that the student has taken the course, 0 indicates that the course was not taken. |
| **CALC** | Enrollment in Calculus; 1 indicates that the student has taken the course, 0 indicates that the course was not taken. |
| **GENDER** | Gender of Respondent; 0 for female, 1 for male. |
| **FAED** | Highest level of father's education, with values ranging from 2 (less than high school) to 10 (Ph.D., M.D., or other advanced degree). |
| **MAED** | Highest level of mother's education, with values ranging from 2 (less than high school) to 10 (Ph.D., M.D., or other advanced degree). |

## SAS Routine

The listing of the SAS routine is found in Table 1; the following is an annotation of the various subsections of the listing.

***Subsection A*** - contains the data definition lines. Note that the data set name given to the original data is **SAMPLE**. (For the specific details on data entry, see the SAS manuals). The data are entered after the CARDS line.

***Subsection B*** - contains a PROC PRINT which is optional since it only provides an "echo check" of the data; the UNIFORM option and the ID OBSERVE lines are just statements for formatting the printout.

***Subsection C*** - contains the lines that select the N observation numbers that will occur in each of the M random samples. These samples are created using only the variable OBSERVE (which corresponds to the ID number in the original data set). The other variables associated with each observation will be merged with each of these random

samples in Subsection D. For this example, we have created 100 samples (M = 100) from the 75 OBSERVE numbers (N = 75). Note that the only tedious part of using this routine is the data entry into the RANTBL function. The integer 100 is the seed for starting the process; the decimal .0133 is simple 1/75 of the reciprocal of the size of the original sample (note that the decimal .0133 is entered 75 times). Any integer can be used as the seed; the remaining decimal entries must be entered judiciously. The KEEP statement is used to output the OBSERVE numbers for each sample into the data set called **SAMPLE**. The last lines in this subsection, PROC FREQ, merely create frequency distributions to demonstrate that each sample has N observations and that each of the N observations contributes uniformly to the total analysis.

***Subsection D*** - contains the lines that are used to merge the original data set (**SAMPLE**) with the data set of bootstrap samples of OBSERVE numbers created in Subsection C (**RANDOM**). Note that prior to the merge state, both data sets must be sorted on the MERGE variable; for this routine we have used OBSERVE as the merge variable. Included in this subsection is a check for errors in the merge process. If there is a complete match as the result of the MERGE statement, then the data set is outputted to a data set called **BOOTSAM** (BOOTstrap SAMple). If there are errors in the process, they are placed in a data set called **ERRORS**. This data set will only be created if there are errors in the implementation of the process.

***Subsection E*** - contains the lines that are used to compute the sample mean(s) from the original data, and then to compute the bootstrap estimate and its standard error. The SAS statement, PROC MEANS, produces simple descriptive statistics for the variables in the VAR statement. In this example, simple statistics will be generated for MATHACH, VISUAL, and MOSAIC. The remaining lines are used to compute the sample means for the variables MATHACH, VISUAL, and MOSAIC for each sample and then compute the mean of these means, the bootstrap estimate ($\hat{\theta}$) and the standard deviation of these means, the bootstrap standard error ($s_{\hat{\theta}}$) The procedure used is PROC MEANS. After the means are computed for each sample, they are outputted to a data

set call **BOOTSTAT** (BOOTstrap STATistics). The final step in the process used PROC UNIVARIATE to compute the bootstrap estimate and standard error (PROC MEANS could also be used).

## Using the SAS Routine

The above SAS routine was used to generate 100, 200 and 500 bootstrap samples and then to compute the bootstrap estimate and its standard error for the following parameters:

- *MEANS* - Variables -- MATHACH, VISUAL and MOSAIC
- *CORRELATION COEFFICIENT* - VISUAL with MOSAIC
- *SIMPLE REGRESSION COEFFICIENT* - Model: MATHACH = VISUAL
- *MULTIPLE REGRESSION COEFFICIENT* - Model: MATHACH = VISUAL MOSAIC
- *DISCRIMINANT FUNCTION COEFFICIENT* - Model: ALG1 = MATHACH
- *FACTOR LOADINGS* - Variables -- MATHACH, VISUAL and MOSAIC

The SAS statements used to compute the bootstrap estimate and their standard error for the means for the variables MATHACH, VISUAL and MOSAIC were discussed above and found in Table 1; the SAS statements for the bootstrap estimates and standard errors for the remaining parameters are also found in Table 1. For each parameter in Table 1, the first step is to compute the descriptive statistic(s) from the original data. The second step is to compute the statistic(s) for each of the M bootstrap samples and to print them. The final step is to compute the bootstrap estimate(s) and their standard errors. Note that for each parameter, other than the factor loadings, the NOPRINT option is used. The only data point(s) of interest is (are) the actual bootstrap statistic(s) for each sample; these statistics are outputted into the data set called **BOOTSTAT** which is printed using PROC PRINT. The NOPRINT option in the procedure statement simply reduces the amount of printout produced. The PROC FACTOR procedure does not have a NOPRINT option.

Note the elaborate SAS statements in PROC CORR, PROC CANDISC and PROC FACTOR which contain a _TYPE_ and/or _NAME_ statement (see SAS Manual for full details). These statements are necessary for storing the bootstrap statistics in the data set called **BOOTSTAT**. For example, consider the output from PROC CORR for the first 5 of the 100 bootstrap samples (see Table 2). The only data point needed is the correlation between VISUAL and MOSAIC; for sample 1, r = -.208446; for sample 2, r = .167087, etc. To KEEP these data points, the following SAS statement must be used.

```
IF _TYPE_ = 'CORR' AND _NAME_ = 'MOSAIC';
   KEEP SAMPLE VISUAL;
```

Likewise, consider the output from the first sample from PROC CANDISC (see Table 3). The two data points needed are the raw score and the standardized discriminant function coefficients. In the output, these data points are RAWSCORE and SCORE, respectively, in the column labelled _TYPE_. To create this data set, the following SAS statement must be used:

```
IF _TYPE_ = 'SCORE' OR _TYPE_ = 'RAWSCORE';
   KEEP SAMPLE _TYPE_ MATHACH;
```

Finally, the SAS statements for PROC FACTOR are similar to those discussed above. In order to create the data set, the following SAS statement must be used:

```
IF _TYPE_ = 'PATTERN' AND _NAME_ = 'FACTOR1';
   KEEP SAMPLE MATHACH VISUAL MOSAIC;
```

## Using the Bootstrap Method

The research literature illustrates two methods for using bootstrap samples. the first is to use the bootstrap samples for estimating parameters ($\hat{\theta}$) and generating their standard errors ($s_{\hat{\theta}}$). For this usage, Efron (1986) indicates that as "few as 10 to 20 bootstrap replications can give quite useful estimates of standard [errors]." However, more than 200 replications are hardly needed for this purpose (Efron, 1986; Efron & Tibshirani, 1986).

For the HSB data set described above, simple descriptive statistics were computed for means (MATHACH, VISUAL, MOSAIC), correlation coefficients (VISUAL with MOSAIC), regression coefficients (MATHACH = VISUAL and MATHACH = VISUAL + MOSAIC), discriminant function coefficients (ALG1 = MATHACH), and factor weights (MATHACH, VISUAL, MOSAIC). These simple statistics are found in Table 4. Note that the standard errors are computed for each statistic except for the discriminant function coefficients and the factor weights; standard errors are not defined for these statistics. Table 5 contains the mean of the bootstrap replications for each statistic as well as their standard errors for M = 100, 200, and 1,000 replications. Note that the standard errors generated from these replications are very consistent and nearly the same as those computed using traditional statistical methods.

The second use of the bootstrap sample involves developing nonparametric confidence intervals (Efron, 1981, 1986; Efron & Tibshirani, 1986; Wu, 1986). For this usage, the number of bootstrap replications is much greater (Efron, 1986; Hall, a986). In fact, Efron (1986) recommends that it is "not excessive to use 2,000 replications," but that 1,000 might be sufficient. In the case of the 1,000 replications, the 90 percent confidence interval would be determined by eliminating the smallest 5 percent of the estimates and the largest 5 percent of the of the estimates. In other words, the endpoints of the 90 percent confidence interval would be the 51st and the 950th ordered statistics of the 1,000 replications. Table 6 contains the 90 percent confidence intervals for each of the bootstrap estimates along with the 90 percent confidence interval developed using traditional statistical methods. For completeness, the 90 percent confidence intervals are given for M = 100, 200 and 1,000 replications. Note again the similarity among the intervals. More importantly, however, note that bootstrap confidence intervals exist for the raw score and standardized discriminant function coefficient and for the factor weights.

The research literature of the past decade has justified the use of bootstrap methods as alternatives to traditional statistical methods which rely on the normality assumption. In fact, it has been shown that the distribution of bootstrap samples "follows the theoretical distributions rather closely (Bickel & Freedman, 1981)." Based upon this literature, researchers have shown the utility of using bootstrap methods for not only simple statistics such as means, medians, correlation and regression coefficients (Efron & Gong, 1986; Lunneborg, 1987), but also discriminant function coefficients (Fenwick, 1979), and factor weights (Chattergee, 1984; Jennrich & Clarkson, 1980). However, even though the Bootstrap method will "undoubtedly remain a major tool in the resampling arsenal (Wu, 1986)," caution must be exercised "against its indiscreet use in complex situations (Rao & Wu, 1985; Schneker, 1985).

**References and Selective Bibliography**


Bickel, P. J. & Freedman, D. A. (1981). Some asymptotic theory for the Bootstrap. *The Annals of Statistics*, 9(6), 1196-1217.

Chattergee, S. (1984). Variance estimation in factor analysis: An application of the bootstrap. *British Journal of Mathematical and Statistical Psychology*, 37, 352-362.

Diaconis, P. & Efron, B. (1983). Computer intensive methods in statistics. *Scientific America*, 248(5), 116-130.

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics*, 7, 1-26.

Efron, B. (1981). Nonparametric estimates of standard error: The Jackknife, the Bootstrap, and other resampling methods. *Biometrika* , 68, 589-599.

Efron, B. (1982). The Jackknife, the Bootstrap and other resampling plans. S.I.A.M., CBMS, National Science Foundation Monograph, 38.

Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika*, 72(1), 45-58.

Efron, B. (1988). Bootstrap confidence intervals: Good or bad? *Psychological Bulletin*, 104(2), 293-296.

Efron, B. & Gong, G. (1986). A leisurely look at the Bootstrap, the Jackknife and cross-validation. *American Statistician*, 37, 36-48.

Efron, B. & Stein, C. (1981). The Jackknife estimate of variance. *The Annals of Statistics*, 9(3) 586-596.

Efron, B. & Tibshirani, R. (1986). "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy." *Statistical Science*, 1(1), 54-57.

Fenwick, I. (1979). Techniques in marketing measurement: The Jackknife. *Journal of Marketing Research*, 16, 410-414.

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, 9(6), 1218-1228.

Hall, P. (1986). In the number of bootstrap simulations required to construct a confidence interval. *The Annals of Statistics*, 14, 1453-1462.

Hinkle, D. E., Wiersma, W. & Jurs, S. G. (1988). *Applied Statistics for the Behavioral Services*. (2nd ed.) Boston: Houghton-Mifflin.

Jennrick, R. I. & Clarkson, D. B. (1980). A feasible method for standard errors of estimate in maximum likelihood factor analysis. *Psychometrika*, 45, 237-247.

Lunneborg, C. E. (1985). Estimating the correlation coefficient: The Bootstrap approach. Psychological Bulletin, 98, 209-215.

Lunneborg, C. E. (1987). Bootstrap applications for the behavioral sciences. *Educational and Psychological Measurements*, 47, 627-629.

Miller, R. G. (1974). The Jackknife - A review. *Biometrika*, 43, 647-649.

Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society*, Series B, 11, 68-84.

Quenouille, M. (1956). Notes on bias in estimation. *Biometrika*, 43, 647-649.

Rasmussen, J. L. (1987). Estimating the correlation coefficient: Bootstrap and parametric approaches. *Psychological Bulletin*, 101, 136-139.

Schenker, N. (1985). Qualms about bootstrap confidence intervals. *Journal of the American Statistical Association*, 80, 360-361

Tukey, J. (1958). Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics*, 29, 614.

Wu, C. F. J. (1986). Jackknife, Bootstrap and other resampling methods in regression analysis. *The Annals of Statistics, 14(4), 1261-1295.*

**Table 1**

**SAS LISTING FOR BOOTSTRAP PROGRAM**

```
X-X-X-X-X-X-X-X-X-X-X-X-X -- SUBSECTION A -- X-X-X-X-X-X-X-X-X-X-X-X-X

DATA SAMPLE;
INPUT  @  1  OBSERVE          3.  /* OBSERVATION NUMBER          */
       @  5  VISUAL           5.2 /*                             */
       @ 11  MOSAIC           4.2 /* FOR OTHER DETAILS SEE THE   */
       @ 16  GRADES           1.  /* LABLES BELOW.               */
       @ 18  MATHGR           1.  /*                             */
       @ 20  ALG1             1.  /*                             */
       @ 22  ALG2             1.  /*                             */
       @ 24  GEO              1.  /*                             */
       @ 26  TRIG             1.  /*                             */
       @ 28  CALC             1.  /*                             */
       @ 31  MATHACH          6.3 /*                             */
       @ 38  FAED             2.  /*                             */
       @ 41  MAED             2.  /*                             */
       @ 44  GENDER           1.  /*                             */
;
LABEL    VISUAL  = 'SCORE ON A VISUALIZATION TEST'
         MOSAIC  = 'SCORE ON PATTERN RELATIONSHIP TEST'
         GRADES  = 'SELF-REPORT OF GRADES'
         MATHGR  = 'SELF-REPORT OF MATHEMATICS GRADES'
         ALG1    = 'ENROLLMENT IN ALGEBRA 1'
         ALG2    = 'ENROLLMENT IN ALGEBRA 2'
         GEO     = 'ENROLLMENT IN GEOMETRY'
         TRIG    = 'ENROLLMENT IN TRIGONOMETRY'
         CALC    = 'ENROLLMENT IN CALCULUS'
         MATHACH = 'SCORE ON A MATHEMATICS ACHIEVEMENT TEST'
         FAED    = 'HIGHEST LEVEL OF EDUCATION OF FATHER'
         MAED    = 'HIGHEST LEVEL OF EDUCATION OF MOTHER'
         GENDER  = 'GENDER OF RESPONDENT'
;
CARDS;
*********************************
   THE DATA ARE TAKEN FROM
HINKLE, WIERSMA & JURS,  1988
*********************************


X-X-X-X-X-X-X-X-X-X-X-X-X -- SUBSECTION B -- X-X-X-X-X-X-X-X-X-X-X-X-X

PROC PRINT UNIFORM DATA=SAMPLE;
ID OBSERVE;
TITLE 'ECHO CHECK OF INPUT DATA';
```

Table 1 (Continued)

```
X-X-X-X-X-X-X-X-X-X-X-X-X -- SUBSECTION C -- X-X-X-X-X-X-X-X-X-X-X-X-X

DATA RANDOM;
M = 100;              /* NUMBER OF SAMPLES                              */
N = 75;               /* NUMBER OF OBSERVATIONS IN EACH SAMPLE          */
UPPER = N;
DO I = 1 TO M;        /* CREATE M RANDOM SAMPLES WITH REPLACEMENT       */
  SAMPLE = I;
  DO J = 1 TO N;      /* SELECT N OBSERVATION NUMBERS (W/REPLACEMENT */
     OBSERVE = .;
     DO WHILE (OBSERVE LT 1 OR OBSERVE GT UPPER);
       OBSERVE = RANTBL(100,.0133,.0133,.0133,.0133,.0133,.0133,.0133,
                            .0133,.0133,.0133,.0133,.0133,.0133,.0133,
                            .0133,.0133,.0133,.0133,.0133,.0133,.0133,
                            .0133,.0133,.0133,.0133,.0133,.0133,.0133,
                            .0133,.0133,.0133,.0133,.0133,.0133,.0133,
                            .0133,.0133,.0133,.0133,.0133,.0133,.0133,
                            .0133,.0133,.0133,.0133,.0133,.0133,.0133,
                            .0133,.0133,.0133,.0133,.0133,.0133,.0133,
                            .0133,.0133,.0133,.0133,.0133,.0133,.0133,
                            .0133,.0133,.0133,.0133,.0133);
     END; /* END DO WHILE */
     KEEP SAMPLE OBSERVE;
     OUTPUT;
  END; /* DO J = 1 TO N */
END; /* DO I = 1 TO M */
TITLE 'GENERATE BOOTSTRAP SAMPLES';

PROC FREQ DATA=RANDOM;
TABLE OBSERVE SAMPLE;
TITLE 'FREQUENCIES FOR OBSERVATION NUMBERS AND OBSERVATIONS PER SAMPLE';

X-X-X-X-X-X-X-X-X-X-X-X-X -- SUBSECTION D -- X-X-X-X-X-X-X-X-X-X-X-X-X

PROC SORT DATA=SAMPLE;
  BY OBSERVE;
PROC SORT DATA=RANDOM;
  BY OBSERVE;
DATA BOOTSAM ERRORS;
MERGE RANDOM (IN = INRAN)
      SAMPLE (IN = INSAM);
BY OBSERVE;
IF (INRAN AND INSAM) THEN OUTPUT BOOTSAM;
                     ELSE OUTPUT ERRORS;
TITLE 'MERGE ORIGINAL DATA WITH BOOTSTRAP SAMPLES';

PROC PRINT DATA=ERRORS;
TITLE 'THESE ARE ERRORS THAT INDICATE A FLAW IN THE ALGORITHM';
```

Table 1 (Continued)

```
X-X-X-X-X-X-X-X-X-X-X-X -- SUBSECTION E -- X-X-X-X-X-X-X-X-X-X-X-X

X-X-X-X-X-X-X-X-X-X-X-X --  MEANS  --  X-X-X-X-X-X-X-X-X-X-X-X

PROC MEANS DATA=SAMPLE;
VAR MATHACH VISUAL MOSAIC;
TITLE 'MEANS FOR THE VARIABLES MATHACH VISUAL MOSAIC';

PROC SORT DATA = BOOTSAM;BY SAMPLE;
PROC MEANS NOPRINT DATA=BOOTSAM;
VAR MATHACH VISUAL MOSAIC;
BY SAMPLE;
OUTPUT  OUT=BOOTSTAT
        MEAN= MATHMEAN VISMEAN MOSMEAN;
PROC PRINT DATA=BOOTSTAT;
ID SAMPLE;
TITLE 'CALCULATE BOOTSTRAP STATISTIC FOR EACH SAMPLE';

PROC UNIVARIATE DATA=BOOTSTAT;
VAR MATHMEAN VISMEAN MOSMEAN;
TITLE 'BOOTSTRAP - STANDARD ERROR OF THE MEAN';

X-X-X-X-X-X-X-X-X-X -- CORRELATION COEFFICIENT -- X-X-X-X-X-X-X-X-X-X

PROC CORR DATA = SAMPLE;
VAR VISUAL MOSAIC;
TITLE 'CORRELATION COEFFICIENT -- VISUAL/MOSAIC';

PROC SORT DATA = BOOTSAM;BY SAMPLE;
PROC CORR NOPRINT OUT=BOOTSTAT DATA=BOOTSAM;
BY SAMPLE;
VAR VISUAL MOSAIC;
DATA BOOTSTAT;SET BOOTSTAT;
IF _TYPE_ = 'CORR' AND _NAME_ = 'MOSAIC';
     KEEP SAMPLE VISUAL;
RENAME VISUAL=CORR;
PROC PRINT DATA=BOOTSTAT;
ID SAMPLE;
TITLE 'COMPUTE CORRELATION - VISUAL/MOSAIC - FOR EACH SAMPLE';

PROC UNIVARIATE DATA=BOOTSTAT;
VAR CORR;
TITLE 'BOOTSTRAP -  STANDARD ERROR OF THE CORR - VISUAL/MOSAIC';
```

Table 1 (Continued)

```
X-X-X-X-X-X-X-X-X-X-X -- SIMPLE REGRESSION -- X-X-X-X-X-X-X-X-X-X-X

PROC REG DATA = SAMPLE;
MODEL MATHACH = VISUAL;
TITLE 'SIMPLE REGRESSION -- MODEL: MATHACH = VISUAL';

PROC SORT DATA = BOOTSAM;BY SAMPLE;
PROC REG NOPRINT DATA=BOOTSAM OUTEST=BOOTSTAT;
MODEL MATHACH = VISUAL;
BY SAMPLE;
PROC PRINT DATA=BOOTSTAT;
ID SAMPLE;
TITLE 'COMPUTE SIMPLE REGRESSION COEFFICIENT FOR EACH SAMPLE';

PROC UNIVARIATE DATA=BOOTSTAT;
VAR VISUAL;
TITLE 'BOOTSTRAP - STANDARD ERROR OF SIMPLE REGRESSION COEFFICIENT';


X-X-X-X-X-X-X-X-X-X-X -- MULTIPLE REGRESSION -- X-X-X-X-X-X-X-X-X-X-X

PROC REG DATA = SAMPLE;
MODEL MATHACH = VISUAL MOSAIC;
TITLE 'MULTIPLE REGRESSION - MODEL: MATHACH = VISUAL MOSAIC';

PROC SORT DATA = BOOTSAM;BY SAMPLE;
PROC REG NOPRINT DATA=BOOTSAM OUTEST=BOOTSTAT;
MODEL MATHACH = VISUAL MOSAIC;
BY SAMPLE;
PROC PRINT DATA=BOOTSTAT;
TITLE 'COMPUTE BOTH REGRESSION COEFFICIENTS FOR EACH SAMPLE';

PROC UNIVARIATE DATA=BOOTSTAT;
VAR VISUAL MOSAIC;
TITLE 'BOOTSTRAP - STANDARD ERROR FOR BOTH REGRESSION COEFFICIENTS';
```

Table 1 (Continued)

```
X-X-X-X-X-X-X-X-X-X-X -- DISCRIMINANT ANALYSIS -- X-X-X-X-X-X-X-X-X-X-X

PROC CANDISC DATA = SAMPLE;
CLASSES ALG1;
VAR MATHACH;
TITLE 'DISCRIMINANT ANALYSIS - MODEL:  ALG1 = MATH';

PROC SORT DATA = BOOTSAM;BY SAMPLE;
PROC CANDISC NOPRINT DATA=BOOTSAM OUTSTAT=BOOTSTAT;
CLASSES  ALG1;
VAR  MATHACH;
BY SAMPLE;
DATA  BOOTSTAT;SET BOOTSTAT;
IF _TYPE_ = 'SCORE' OR _TYPE_ = 'RAWSCORE';
     KEEP SAMPLE _TYPE_ MATHACH;
PROC PRINT DATA=BOOTSTAT;
ID SAMPLE;
TITLE 'COMPUTE DISCRIMINANT COEFFICIENT FOR EACH SAMPLE';

PROC SORT;BY _TYPE_;
PROC UNIVARIATE DATA=BOOTSTAT;
BY _TYPE_;
VAR MATHACH;
TITLE 'BOOTSTRAP - STANDARD ERROR OF THE DISCRIMINANT COEFFICIENT';


X-X-X-X-X-X-X-X-X-X-X -- FACTOR ANALYSIS -- X-X-X-X-X-X-X-X-X-X-X

PROC FACTOR DATA = SAMPLE;
VAR MATHACH VISUAL MOSAIC;
TITLE 'FACTOR ANALYSIS -- VARIABLES -- MATHACH VISUAL MOSAIC';

PROC SORT DATA = BOOTSAM;BY SAMPLE;
PROC FACTOR OUTSTAT = BOOTSTAT DATA = BOOTSAM;
VAR MATHACH VISUAL MOSAIC;
BY SAMPLE;
DATA BOOTSTAT;SET BOOTSTAT;
IF _TYPE_ = 'PATTERN' AND _NAME_ = 'FACTOR1';
     KEEP SAMPLE MATHACH VISUAL MOSAIC;
PROC PRINT DATA=BOOTSTAT;
ID SAMPLE;
TITLE 'COMPUTE FACTOR WEIGHTS FOR EACH SAMPLE';

PROC UNIVARIATE DATA=BOOTSTAT;
VAR MATHACH VISUAL MOSAIC;
TITLE 'BOOTSTRAP -- STANDARD ERROR OF THE FACTOR WEIGHTS';
```

Table 2

SAS OUTPUT FROM *PROC CORR*

| SAMPLE | _TYPE_ | _NAME_ | VISUAL | MOSAIC |
|---|---|---|---|---|
| 1 | MEAN | | 5.76333 | 27.8133 |
| 1 | STD | | 4.14291 | 8.73281 |
| 1 | N | | 75 | 75 |
| 1 | CORR | VISUAL | 1 | -.208446 |
| 1 | CORR | MOSAIC | -.208446 | 1 |
| 2 | MEAN | | 4.83 | 26.86 |
| 2 | STD | | 3.62405 | 8.01374 |
| 2 | N | | 75 | 75 |
| 2 | CORR | VISUAL | 1 | 0.167087 |
| 2 | CORR | MOSAIC | 0.167087 | 1 |
| 3 | MEAN | | 4.86667 | 27.1133 |
| 3 | STD | | 3.62642 | 9.87988 |
| 3 | N | | 75 | 75 |
| 3 | CORR | VISUAL | 1 | 0.107874 |
| 3 | CORR | MOSAIC | 0.107874 | 1 |
| 4 | MEAN | | 4.42333 | 28.26 |
| 4 | STD | | 4.00548 | 9.59496 |
| 4 | N | | 75 | 75 |
| 4 | CORR | VISUAL | 1 | 0.120559 |
| 4 | CORR | MOSAIC | 0.120559 | 1 |
| 5 | MEAN | | 6.07333 | 27.4 |
| 5 | STD | | 3.73641 | 9.63769 |
| 5 | N | | 75 | 75 |
| 5 | CORR | VISUAL | 1 | -0.06964 |
| 5 | CORR | MOSAIC | -0.06964 | 1 |

Table 3

SAS OUTPUT FROM *PROC CANDISC*

| SAMPLE | ALG1 | _TYPE_ | _NAME_ | MATHACH |
|---|---|---|---|---|
| 1 | . | SSCP | MATHACH | 2751.79 |
| 1 | . | COV | MATHACH | 37.19 |
| 1 | . | BSSCP | MATHACH | 443.87 |
| 1 | . | WSSCP | MATHACH | 2307.92 |
| 1 | . | BCOV | MATHACH | 11.84 |
| 1 | . | WCOV | MATHACH | 31.62 |
| 1 | . | CORR | MATHACH | 1.00 |
| 1 | . | BCORR | MATHACH | 1.00 |
| 1 | . | WCORR | MATHACH | 1.00 |
| 1 | . | MEAN | | 14.13 |
| 1 | . | STD | | 6.10 |
| 1 | . | BSTD | | 3.44 |
| 1 | . | WSTD | | 5.62 |
| 1 | . | RSQUARED | | 0.16 |
| 1 | . | VARRATIO | | 0.19 |
| 1 | . | F | | 14.04 |
| 1 | . | PROBF | | 0.00 |
| 1 | . | N | | 75.00 |
| 1 | 0 | N | | 13.00 |
| 1 | 1 | N | | 62.00 |
| 1 | 0 | MEAN | | 8.82 |
| 1 | 1 | MEAN | | 15.25 |
| 1 | 0 | TSTDMEAN | | -0.87 |
| 1 | 1 | TSTDMEAN | | 0.18 |
| 1 | 0 | WSTDMEAN | | -0.94 |
| 1 | 1 | WSTDMEAN | | 0.20 |
| 1 | . | CANCORR | | 0.40 |
| 1 | . | TSTRUCT | CAN1 | 1.00 |
| 1 | . | BSTRUCT | CAN1 | 1.00 |
| 1 | . | WSTRUCT | CAN1 | 1.00 |
| 1 | . | SCORE | CAN1 | 1.08 |
| 1 | . | RAWSCORE | CAN1 | 0.18 |
| 1 | 0 | CANMEAN | | -0.94 |
| 1 | 1 | CANMEAN | | 0.20 |

**Table 4**

**TRADITIONAL STATISTICS**
**(N = 75)**

| | $\hat{\theta}$ | $s_{\hat{\theta}}$ |
|---|---|---|
| **MEANS** | | |
| MATHACH | 12.5645 | 0.7702 |
| VISUAL | 5.2433 | 0.4517 |
| MOSAIC | 27.4133 | 1.1055 |
| **CORRELATION COEFFICIENT (VISUAL with MOSAIC)** | | |
| CORR | .0301 | 0.1170 |
| **REGRESSION COEFFICIENT (MATHACH = VISUAL)** | | |
| VISUAL | 0.7214 | 0.1808 |
| **REGRESSION COEFFICIENT (MATHACH = VISUAL MOSAIC)** | | |
| VISUAL | 0.7112 | 0.1776 |
| MOSAIC | 0.1397 | 0.0726 |
| **DISCRIMINANT FUNCTION COEFFICIENT (ALG 1 = MATHACH)** | | |
| RAWSCORE | 0.1697 | (No Standard Errors) |
| STANDARDIZED | 1.1321 | |
| **FACTOR WEIGHTS** | | |
| MATHACH | 0.8504 | |
| VISUAL | 0.7660 | (No Standard Errors) |
| MOSAIC | 0.4199 | |

Table 5

BOOTSTRAP SAMPLES

| | M = 100 | | M = 200 | | M = 1000 | |
|---|---|---|---|---|---|---|
| | $\hat{\theta}$ | $s_{\hat{\theta}}$ | $\hat{\theta}$ | $s_{\hat{\theta}}$ | $\hat{\theta}$ | $s_{\hat{\theta}}$ |
| **MEANS** | | | | | | |
| MATHACH | 12.6493 | 0.7808 | 12.6440 | 0.7405 | 12.5714 | 0.7764 |
| VISUAL | 5.3282 | 0.4501 | 5.2801 | 0.4438 | 5.2680 | 0.4552 |
| MOSAIC | 27.5035 | 1.1031 | 27.4815 | 1.1189 | 27.4144 | 1.1358 |
| **CORRELATION COEFFICIENT (VISUAL with MOSAIC)** | | | | | | |
| CORR | 0.0202 | 0.1015 | 0.0302 | 0.0928 | 0.0334 | 0.0966 |
| **REGRESSION COEFFICIENT (MATHACH = VISUAL)** | | | | | | |
| VISUAL | 0.7218 | 0.1826 | 0.7356 | 0.1818 | 0.7281 | 0.1711 |
| **REGRESSION COEFFICIENT (MATHACH = VISUAL MOSAIC)** | | | | | | |
| VISUAL | 0.7099 | 0.1701 | 0.7208 | 0.1702 | 0.7124 | 0.1622 |
| MOSAIC | 0.1548 | 0.0839 | 0.1554 | 0.723 | 0.1462 | 0.0735 |
| **DISCRIMINANT FUNCTION COEFFICIENT (ALG 1 = MATHACH)** | | | | | | |
| RAWSCORE | 0.1741 | 0.0124 | 0.1725 | 0.0135 | 0.1724 | 0.0123 |
| STANDARDIZED | 1.1469 | 0.0628 | 1.1425 | 0.0626 | 1.1368 | 0.0594 |
| **FACTOR WEIGHTS** | | | | | | |
| MATHACH | 0.8438 | 0.0479 | 0.8469 | 0.0404 | 0.8449 | 0.0429 |
| VISUAL | 0.7636 | 0.0933 | 0.7600 | 0.0849 | 0.7629 | 0.0787 |
| MOSAIC | 0.3567 | 0.2754 | 0.3899 | 0.2447 | 0.3834 | 0.2451 |

**Table 6**

**BOOTSTRAP CONFIDENCE INTERVALS**

| | Parametric Intervals | M = 100 | M = 200 | M = 1000 |
|---|---|---|---|---|
| **MEANS** | | | | |
| MATHACH | (11.0549,14.0741) | (11.4534,14.0044) | (11.4800,13.8622) | (11.3111,13.8623) |
| VISUAL | (4.3580,6.1286) | (4.5867,6.1200) | (4.5467,6.0867) | (4.5133,6.0067) |
| MOSAIC | (25.2465,29.5801) | (25.8267,29.4600) | (25.7800,29.4533) | (25.6267,29.3533) |
| **CORRELATION COEFFICIENT (VISUAL with MOSAIC)** | | | | |
| CORR | (-0.1992,0.2594) | (-0.1384,0.1671) | (-0.1276,0.1795) | (-0.1264,0.1893) |
| **REGRESSION COEFFICIENT (MATHACH = VISUAL)** | | | | |
| VISUAL | (0.3670,1.0758) | (0.4635,1.0157) | (0.4256,1.0157) | (0.4565,1.0089) |
| **REGRESSION COEFFICIENT (MATHACH = VISUAL MOSAIC)** | | | | |
| VISUAL | (0.3631,1.0593) | (0.4682,1.0274) | (0.4231,1.0274) | (0.4473,0.9755) |
| MOSAIC | (-0.0026,0.2820) | (0.0372,0.3119) | (0.0490,0.2674) | (0.0365,0.2685) |
| **DISCRIMINANT FUNCTION COEFFICIENT (ALG 1 = MATHACH)** | | | | |
| RAWSCORE | na | (0.1551,0.1958) | (0.1535,0.0.1961) | (0.1539,0.1932) |
| STANDARDIZED | na | (1.0474,1.2581) | (1.0502,1.1.2554) | (1.0500,1.2437) |
| **FACTOR WEIGHTS** | | | | |
| MATHACH | na | (0.7330,0.8866) | (0.7903,0.8859) | (0.7877,0.8876) |
| VISUAL | na | (0.6515,0.8617) | (0.6304,0.8555) | (0.6315,0.8591) |
| MOSAIC | na | (-0.2829,0.6545) | (-0.0463,0.6581) | (-0.1240,0.6788) |